IT IS CLAIMED:

1. A computer-executed method for matching a target document in the form of a digitally encoded natural-language text with a plurality of sample texts, comprising the steps of:

(a) for each of a plurality of terms selected from one of (i) non-generic words in the document, (ii) proximately arranged word groups in the document, and (iii) a combination of (i) and (ii), determining a selectivity value calculated as the frequency of occurrence of that term in a library of texts in one field, relative to the frequency of occurrence of the same term in one or more other libraries of texts in one or more other fields, respectively, and

(b) representing the document as a vector of terms, where the coefficient assigned to each term is a function of the selectivity value determined for that term,

(c) determining for each of a plurality of sample texts, a match score related to the number of terms present in or derived from that text that match those in the target document, and

(d) selecting one or more of the sample texts having the highest match scores.

2. The method of claim 1, wherein the sample texts are texts in the libraries of texts from which the selectivity values of target terms are determined.

3. The method of claim 1, wherein the selectivity value associated with a term in the document is related to the greatest selectivity value determined with respect to each of a plurality $N \geq 2$ of libraries of texts in different fields.

4. The method of claim 1, wherein the selectivity value assigned to a term is a root function of the frequency of occurrence of that term in said library, relative to the frequency of occurrence of the same term in one or more other libraries of texts in one or more other fields, respectively, and the match score is weighted by

the selectivity values of the matching terms. .

5. The method of claim 4, wherein the root function is between 2, the square root function, and 3, the cube root function.

6. The method of claim 1, wherein only terms having a selectivity value above a predetermined threshold are included in the vector.

7. The method of claim 1, wherein the terms include words in the document, and the coefficient assigned to each word in the vector is also related to the inverse document frequency of that word in one or more of said libraries of texts.

8. The method of claim 6, wherein the coefficient assigned to each word in the vector is the product of a function of the selectivity value and the inverse document frequency of that word.

9. The method of claim 1, wherein the terms include words in the document, and step (a) includes (i) accessing a database of word records, where each record includes text identifiers of the library texts that contain that word, and associated library identifiers for each text, and (ii) using the identified text and library identifiers to calculate one or more selectivity values for that word.

10. The method of claim 9, wherein carrying out the step of determining match scores includes (i) accessing said database of word records to identify library texts associated with each descriptive word in the target text, and (ii) from the identified texts recorded in step (i), determining text match score based on the number of descriptive words in that text weighted by the selectivity values of the matching words.

11. The method of claim 1, wherein the terms include word groups in the document, and said database further includes, for each word record, word-position

identifiers, and wherein step (a) as applied to word groups includes (i) accessing said database to identify texts and associated library and word-position identifiers associated with that word group, (ii) from the identified texts, library identifiers, and word-position identifiers recorded in step and (i) determining one or more selectivity values for that word group.

12. The method of claim 11, wherein carrying out the step of determining match scores includes (i) recording the texts associated with each descriptive word group, and (ii) determining a text match score based, at least in part, on number of descriptive word groups in a text, weighted by the selectivity values of such words groups.

13. An automated system for matching a target document in the form of a digitally encoded natural-language text with a plurality of sample texts, comprising

(1) a computer,

(2) accessible by said computer, a database of word records, where each record includes text identifiers of the library texts that contain that word, associated library identifiers for each text, and optionally, one or more selectivity values for each word, where the selectivity value of a term in a library of texts in a field is related to the frequency of occurrence of that term in said library, relative to the frequency of occurrence of the same term in one or more other libraries of texts in one or more other fields, respectively,

(3) a computer readable code which is operable, under the control of said computer, to perform steps comprising:

(a) for each of a plurality of terms selected from one of (i) non-generic words in the document, (ii) proximately arranged word groups in the document, and (iii) a combination of (i) and (ii), determining a selectivity value calculated as the frequency of occurrence of that term in a library of texts in one field, relative to the frequency of occurrence of the same term in one or more other libraries of texts in one or more other fields, respectively, and

(b) representing the document as a vector of terms, where the coefficient assigned to each term is a function of the selectivity value determined for that

term,

(c) determining for each of a plurality of sample texts, a match score related to the number of vector terms present in or derived from that text that match those in the target document, and

(d) selecting one or more of the sample texts having the highest match scores.

14. The system of claim 13, wherein the terms include words in the document, and step (a) includes (i) accessing a database of word records, where each record includes text identifiers of the library texts that contain that word, and associated library identifiers for each text, and (ii) using the identified text and library identifiers to calculate one or more selectivity values for that word.

15. The system of claim 14, wherein carrying out the step of determining match scores includes (i) accessing said database of word records to identify library texts associated with each descriptive word in the target text, and (ii) from the identified texts recorded in step (i), determining text match score based on the number of descriptive words in that text weighted by the selectivity values of the matching words.

16. The system of claim 13, wherein the terms include word groups in the document, and said database further includes, for each word record, word-position identifiers, and wherein step (a) as applied to word groups includes (i) accessing said database to identify texts and associated library and word-position identifiers associated with that word group, (ii) from the identified texts, library identifiers, and word-position identifiers recorded in step and (i) determining one or more selectivity values for that word group.

17. The system of claim 16, wherein carrying out the step of determining match scores includes (i) recording the texts associated with each descriptive word group, and (ii) determining a text match score based, at least in part, on number of descriptive word groups in a text, weighted by the selectivity values of such words

groups.

18. Computer readable code for use with an electronic computer and a database word records for matching a target document in the form of a digitally encoded natural-language text with a plurality of sample texts, where each record in the word records database includes text identifiers of the library texts that contain that word, an associated library identifier for each text, and optionally, one or more selectivity values for each word, where the selectivity value of a term in a library of texts in a field is related to the frequency of occurrence of that term in said library, relative to the frequency of occurrence of the same term in one or more other libraries of texts in one or more other fields, respectively, said code being operable, under the control of said computer, to perform steps comprising:

(a) for each of a plurality of terms selected from one of (i) non-generic words in the document, (ii) proximately arranged word groups in the document, and (iii) a combination of (i) and (ii), determining a selectivity value calculated as the frequency of occurrence of that term in a library of texts in one field, relative to the frequency of occurrence of the same term in one or more other libraries of texts in one or more other fields, respectively, and

(b) representing the document as a vector of terms, where the coefficient assigned to each term is a function of the selectivity value determined for that term.

(c) determining for each of a plurality of sample texts, a match score related to the number of vector terms present in or derived from that text that match those in the target document, and

(d) selecting one or more of the sample texts having the highest match scores.